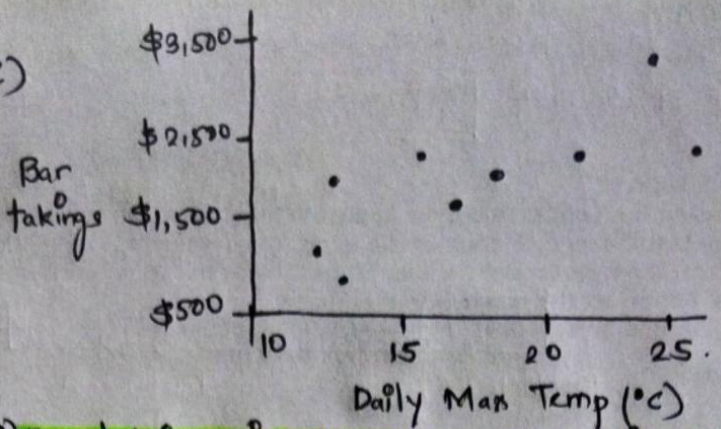
Sum of Square (SS)

Day	Takings	Temp (°C)
3 Jun	$3,213	28
10 Jun	$2,089	21
17 Jun	$2,253	25
21 Jun	$1,801	18
1 Jun	$801	13
8 Jun	$1,991	16
15 Jun	$1,720	13
22 Jun	$1,514	17
29 Jun	$1,017	12

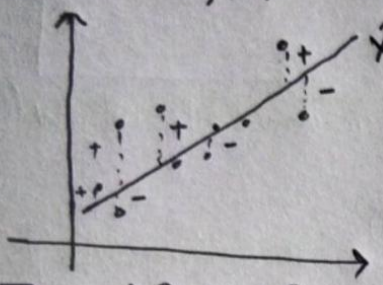


Q) Bar taking given the temperature for the particular day is the dataset.

→ $\hat{y}$ is the predicted value for given value of x.

So, $\hat{y}$ will have a equation, $\hat{y} = -353.11 + 123.54x$

↑ constant term / y intercept ↑ Gradient / Coefficient of x

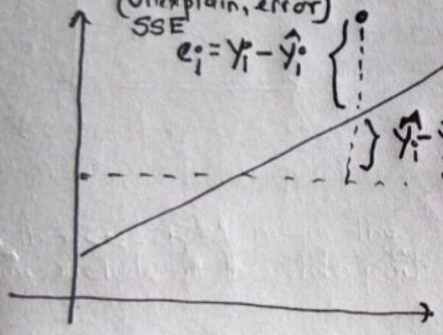


There are 2 kind of error terms - i) +ve error ii) -ve error

But if we sum both +ve and -ve, we will get 0.

So we square, to ignore negative values.

There $\hat{y}$ is the line which decreases the sum of square error (SSE).



Explained component, $\hat{y}_i - \bar{y}_i$

Unexplained component, $y_i - \hat{y}_i$

Total variation = Explained variation + unexplained variation

SST = SSR + SSE

Total deviation = SSR + SSE

SST → Sum of square total

SSR → Sum of square residual = $\hat{y}_i - \bar{y}$

SSE → Sum of square error = $y_i - \hat{y}_i$

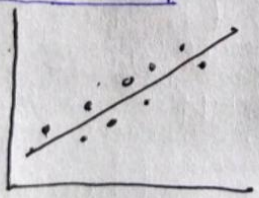
$= \hat{y}_i - \bar{y} + y_i - \hat{y}_i$

$SST = \Sigma(y_i - \bar{y})^2$

$R^2 = SSR/SST$

R^2 is proportion of total variation which is explained.

High SSE
Low R^2
(Scattered point)
Huge error term



Low SSE
High R^2
(Unscattered point)
Low error term

- Regression line is a line which best fit to the observations.

$R^2 = SSR/SST$ = The proportion of the variation in y being explained by the variation in x. R square range from 0 to 1.

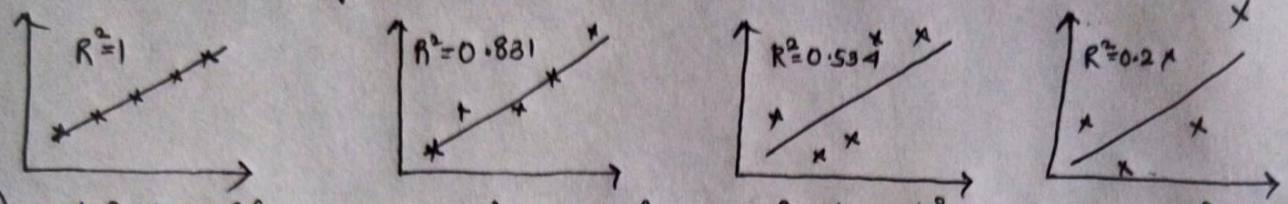
$SSR + SSE = SST$

Sum of square due to regression | Sum of square due to error.

Also known as,

$ESS + RSS = SST/TSS$

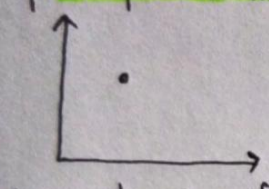
Explained Sum of square | Residual Sum of square | Total Sum of Square.



Q) What is the minimum number of observations require to estimate the regression?

$Y_i = B_0 + B_1 X_i + \varepsilon_i$.
(Height) (Weight)

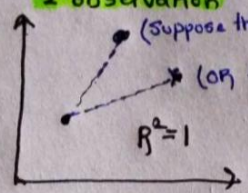
If one point is there



we cannot make a line, so we need two point

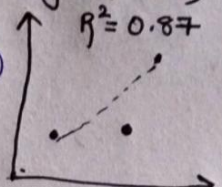
In this case, $R^2=1$ means the line include all the data point

2 observation



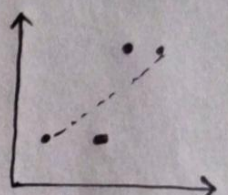
So whenever second point is, line will join. But we cannot check strength of /test of relationship.

3rd observation (In order to check strength of line)



So, 1 point will check strength of line.
So, df = 1
(degree of freedom)

4 observations



$R^2 = 0.772$
$df = 2$.
$df = 4-1-1$
$df = 2$

$Y_i = B_0 + B_1 X_{1i} + B_2 X_{2i} + \varepsilon_i$. How many minimum number of observation require to construct regression line?
(height) (Mothers height)

$df = 5-2-1 = 5-3 = 2$.

- We need minimum 3 points to construct a plane.

$N=3, R^2=1$. $N=4, R^2=0.80, df=1$. $N=5, R^2=0.73, df=2$.

- So, degree of freedom, $df = n - k - 1$

n = number of observation
k = number of explanatory (x) variable (Independent)

Q) How does degree of freedom related to R square?

→ As degree of freedom (df) decreases (i.e, more variable added to given model) R square will only increase

So, even if we add useless variable, R square will only increase. So we use Adjusted R^2.

Adjusted R^2 $\quad$ Adjusted $R^2 = (1) - (1-R^2)\frac{n-1}{n-k-1}$ or $1 - \frac{SSE}{SST}\left(\frac{n-1}{n-k-1}\right)$

as k increases, Adjusted R^2 will tend to decrease, reflecting the reduced power in the model. Only if we add useful variable to the model, Adjusted R^2 will only increase.

Number of observation (n)	number of variable (k)	R^2	Adj-R^2	
25	4	0.71	0.65	
25	5	0.76	0.69	
25	6	0.78	0.71	→ Choose this one.
25	7	0.79	0.70	